Validation of a Virtual Sound Environment System for Testing Hearing Aids

Cubick, Jens; Dau, Torsten

Published in:
Acta Acustica united with Acustica

Link to article, DOI:
10.3813/AAA.918972

Publication date:
2016

Document Version
Peer reviewed version

Validation of a Virtual Sound Environment System for Testing Hearing Aids

J. Cubick, T. Dau
Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, Ørsteds Plads, building 352, 2800 Kgs. Lyngby, Denmark. jecu@elektro.dtu.dk

Summary
In the development process of modern hearing aids, test scenarios that reproduce natural acoustic scenes have become increasingly important in recent years for the evaluation of new signal processing algorithms. To achieve high ecological validity, such scenarios should include components like reverberation, background noise, and multiple interfering talkers. Loudspeaker-based sound field reproduction techniques, such as higher-order Ambisonics, allow for the simulation of such complex sound environments and can be used for realistic listening experiments with hearing aids. However, to successfully employ such systems, it is crucial to know how experimental results from a virtual environment translate to the corresponding real environment. In this study, speech reception thresholds (SRTs) were measured with normal-hearing listeners wearing hearing aids, both in a real room and in a simulation of that room auralized via a spherical array of 29 loudspeakers, using either Ambisonics or a nearest loudspeaker method. The benefit from a static beamforming algorithm was considered in comparison to a hearing aid setting with omnidirectional microphones. The measured SRTs were about 2-4 dB higher, and the benefit from the beamformer setting was, on average, about 1.5 dB smaller in the virtual room than in the real room. These differences resulted from a more diffuse sound field in the virtual room as indicated by differences in measured directivity patterns for the hearing aids and interaural cross-correlation coefficients. Overall, the considered VSE system may represent a valuable tool for testing the effects of hearing-aid signal processing on physical and behavioural outcome measures in realistic acoustic environments.

PACS no. 43.55.Hy, 43.55.Ka, 43.66.Ts

1. Introduction

Hearing aid (HA) users often have difficulties following a conversation in challenging listening situations that involve multiple talkers, background noise and/or reverberation [1], even though they typically benefit from their HAs in simple acoustic situations, such as a one-to-one conversation in a quiet room. The processing power of HAs has increased dramatically over the last 10 years and advanced signal processing strategies have been applied to help the users, particularly in complex listening situations. To assess and evaluate the performance of modern HAs, the test scenarios should therefore be as realistic as possible. Until recently, however, most testing has been done either in very basic conditions with simple loudspeaker setups in acoustically dampened rooms, or in field studies where the end users wear certain types of HAs for some time and report back via questionnaires after the testing period. The first approach offers much control over the test conditions but provides only very limited flexibility regarding the acoustic conditions and does therefore not reflect the challenges that HA users face in their everyday life. In field tests, representing the second approach, the participants experience the HAs in the environments where they would actually use them but the experimental conditions are difficult to control. The simulation of realistic acoustic scenes under controlled and repeatable conditions in the laboratory would combine the advantages of the two approaches.

One well-known method to provide such simulated scenes are headphone-based reproduction systems that use binaural technology [2] to reproduce the correct sound pressure at the listeners' ear. However, even though the results obtained with this method can be very convincing, headphone-based systems have some disadvantages. The simulation is most convincing if it is based on head-related transfer functions that are measured for each listener indi-

Received 15 October 2015,
accepted 13 March 2016.

vidually, and if head tracking is used to keep the auditory image position stable, even if the listener moves his/her head. Measuring impulse responses for all incidence angles requires an enormous measuring effort and makes testing difficult. Furthermore, using HAs under headphones is impractical, as the acoustics under earphone cups are very different from a free field situation. These problems can be avoided with loudspeaker-based technologies that try to reproduce a desired sound field in a room. Sound field reproduction techniques, like wave-field synthesis [3], higher-order Ambisonics (HOA) [4, 5, 6], directional audio coding [7], or direct mapping of reflections to the nearest loudspeaker [8], make it possible to render realistic and reproducible virtual sound environments (VSEs) in the laboratory, including room reverberation and multiple sound sources. In the case of HOA, the system aims at reproducing the sound field correctly at the listener's location in the virtual room around the "sweet spot" in the centre of the loudspeaker array. The presence of the listener thus ideally generates exactly the same acoustic effects as it would in the real sound field. Head rotations are allowed and, unlike in headphone-based systems, listeners are able to wear HAs in a VSE. In a HOA-based system, however, the spatial resolution of the reproduced sound field is limited by the Ambisonics order which, in turn, depends on the number of loudspeakers in the array [5].

Such a HOA-based system has been realized at the Technical University of Denmark. It comprises a spherical array of 29 loudspeakers mounted in an acoustically highly dampened room (see Figure 1). The VSEs are based on simulations using the room acoustic modelling software ODEON [9]. A 3-dimensional model of a room is generated and the absorption and scattering properties of all surfaces are defined, as well as all source positions and the receiver position and direction. Even though such a geometrical acoustics-based simulation has limitations, especially in the low frequencies and with small rooms, it is very easy to model very well-defined complex listening scenarios. The simulation results are then processed by the loudspeaker-based room auralization (LoRA) toolbox [10]. Using either HOA or a method where each reflection is mapped to the nearest loudspeaker (NLS), a multi-channel room impulse response is generated, which, when convolved with an anechoic source signal, yields the driving signal for the loudspeakers. Several studies have been conducted to evaluate the performance of this system. One study compared the common room acoustic parameters, defined in [11] and derived from the LoRA output, with the corresponding values provided by the underlying ODEON simulation [10]. Considering different seats in a classroom and a concert hall, it was found that the variation of the room acoustic parameters for small head movements was mostly within 1-2 difference limens [12, 13] of the ODEON results. In another study [14], speech intelligibility in noise was measured for different rendering methods. The highest speech intelligibility was found when NLS coding was used, whereas it was lower in the case of 4^{th}-order HOA and even lower in the case of 1^{st}-order Ambisonics. In a third study [15], distance perception in the VSE was investigated and no significant difference was found between the LoRA system and a test based on binaural recordings. A study with a technically comparable auralization system at the HA manufacturer Oticon [16] compared speech intelligibility and listening effort of hearing-impaired listeners in different virtual rooms, a 'dry' room, a lecture hall, and a very reverberant basement. Another study, using a similar system, tested speech intelligibility in a 'complex' cafeteria environment with multiple talkers, and in a 'standard' anechoic environment [17]. Finally, two very recent simulation studies investigated the applicability of multichannel loudspeaker-based reproduction chains for testing HAs [18, 19].

Figure 1: Photograph of the 'Spacelab' at DTU. A spherical array of 29 loudspeakers allows for the auralization of acoustical scenes in virtual rooms. Photo: Joachim Rode.

However, in all above studies, the VSE systems were evaluated either by comparing theoretical quantities, or room acoustical measures between the VSE and the underlying ODEON simulation, or by comparing results of behavioural measurements obtained inside the system. Only a few studies actually compared the listening performance measured in a VSE with the performance in the corresponding real environment. Few studies used simulation-based auralizations presented via headphones and compared speech intelligibility in this setup with the one measured in the real rooms, e.g. [20, 21, 22], or overall listening experience [23]. One early study compared speech intelligibility in a loudspeaker-based auralization system and in a real room using binaural

technology [24], and, to the knowledge of the authors, only one study has compared perceptual measures obtained in a loudspeaker-based VSE directly to the same measures obtained in the corresponding real room [25]. To successfully employ the system for HA testing, it is crucial to know how well experimental results from a VSE translate to real-life situations.

Specifically, the present study investigated whether the reproduction of a VSE in the LoRA-based system captures the acoustic properties of a 40-seat classroom accurately enough, such that the effects of HA processing in the VSE can be considered to be the same as, or very close to, the real environment. To achieve this goal, three requirements need to be fulfilled: (1) The ODEON simulation must be well calibrated to capture the key acoustical properties of the classroom. To assure this, the simulation results for the common room acoustic parameters reverberation time, T_{30}, and clarity for speech, C_{50}, [11] from ODEON were compared to the values measured in the classroom; (2) The LoRA processing must be transparent to preserve these properties. To test the transparency of the LoRA processing, the same room acoustic parameters were calculated from room impulse responses measured inside the VSE, using either HOA or NLS rendering; and (3) The HA performance in the VSE and the real room needs to be comparable. To assess the HA performance, directivity patterns were measured both in the classroom and the VSE, using omnidirectional microphones and a static beamforming (BF) program [26].

If these requirements are fulfilled, the performance of the listeners in behavioural tasks in the VSE and the real room may be assumed to be comparable. To evaluate this, speech intelligibility was considered as an outcome measure in the present study since it represents one of the most important performance indicators in the HA development process. Speech reception thresholds (SRTs) were measured both in the classroom and its virtual counterpart with normal-hearing listeners, either with or without HAs. Testing normal-hearing listeners with HAs might seem counterintuitive but was chosen here as a first step in the evaluation process of the VSE system; normal-hearing listeners typically show more "homogeneous" results than hearing-impaired listeners and the main focus of the present study was to study the effect of basic features in the HA settings on the selected outcome measures in the real versus the simulated environments. The SRT benefit from a static BF algorithm relative to a HA setting with omnidirectional microphones was tested. This algorithm has been shown to yield a speech perception benefit of up to 8.5 dB in optimized conditions, when the test was performed in a sound-insulated booth with noise presented from 180° azimuth, [27], or up to 3.9 dB in more realistic scenarios with a noise source at 90° azimuth in a room with a reverberation time of 0.45 s [28].

It was hypothesized in the present study that inaccuracies in the sound field reproduction should decrease the effectiveness of the BF and the associated gain in the effective signal-to-noise ratio (SNR) for frontal sources, which should result in higher SRTs. It was assumed that the room simulation can be considered sufficiently authentic if (1) the SRTs measured in the VSE are close to those obtained in the corresponding real room and if (2) threshold differences between the two HA settings are similar in the two situations.

2. Methods

2.1. Auralization technique

The acoustical data for the VSEs in the system under test were generated based on a room simulation in the commercial room acoustic simulation software ODEON [9]. This software uses a hybrid method for the calculation of the room acoustic parameters [29, 30]. The image source and ray tracing methods [6] are combined to calculate the reflections up to a certain order. Above this transition order, the secondary source method is used to compute the late part of the room impulse response (RIR). The ODEON simulations in this study were run with 8000 early rays, 8000 late rays, a maximum reflection order of 2000, an impulse response resolution of 1 ms and a transition order of 3. The virtual sound sources were modelled to have the same directivity in the horizontal plane as that measured in an anechoic chamber for the Dynaudio BM6P loudspeaker used as the target source in the listening experiments. The simulation results, i.e., the reflectogram, containing information about the delay, direction and frequency content of each early reflection up to the transition order, and the energy decay curves, were exported from ODEON and processed by the LoRA toolbox [10] to generate the driving signals for the loudspeaker array.

Due to the precedence effect [31, 32], the localization of a sound source in a room is mostly determined by the direct sound, whereas the late reflections in the rather diffuse reverberant tail of the RIR cannot be resolved individually [33]. Following these properties of human sound localization, the LoRA toolbox splits the RIR into the direct sound, the early reflections, and the late reflections. The direct sound and the early reflections up to the transition order are rendered with the highest possible resolution, i.e., by either employing the highest possible HOA order for a given loudspeaker array, or by mapping it to the nearest loudspeaker available (NLS). The late reflections are provided by ODEON as the vectorial intensity and

the envelope of the energy. These data are interpreted as a 1st order Ambisonics signal and are decoded correspondingly. The resulting envelope for the late reflections is then multiplied with uncorrelated noise for each loudspeaker [10]. Summing up the parts of the decoded RIR generates a multi-channel RIR, and convolution of this RIR with an anechoic signal forms the driving signal for the loudspeakers.

The VSE in the listening tests was played back through the spherical array of 29 Dynaudio BM6P loudspeakers in the 'Spacelab' shown in Figure 1. The array consists of a horizontal ring of 16 loudspeakers at ear height of a sitting listener at a distance of 1.8 m, two rings of 6 loudspeakers at ±45° elevation and one loudspeaker on the ceiling above the centre of the array. It is placed in an acoustically dampened room with a reverberation time of 0.16 s in the 125-Hz octave band and below 0.1 s in all frequency bands above 125 Hz. All loudspeakers were equalized to a flat frequency response relative to an omni-directional B&K 4092 microphone in the centre of the array using 1114-tap FIR filters. In the listening tests, 4th order three-dimensional HOA rendering was used.

The room chosen for the VSE in this study was "Room 019", a lecture room at DTU with 40 seats and a volume of about 180 m^3. The ODEON model was carefully matched to the reverberation time and clarity values measured at the listening position shown in Figure 2 by assigning materials with appropriate absorption and scattering coefficients to the model surfaces. In addition to T_{30}, Clarity was considered an important criterion for the model calibration, because this early-to-late energy ratio is related to speech intelligibility [11].

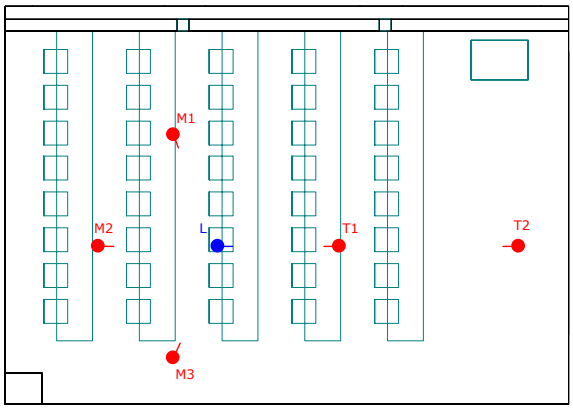


Figure 2: Top view of the room model with the listening position (L), the three maskers (M1, M2, M3), and the target speech sources T1 at 2 m and T2 at 5 m.

2.2. Physical evaluation

2.2.1. Room acoustic parameters

For the physical validation of the VSE, the common room acoustic parameters reverberation time, T_{30}, clarity for speech, C_{50}, and the interaural cross-correlation coefficient, $IACC$, were calculated according to [11] from RIRs measured with logarithmic sine sweeps [34]. This was done both in the classroom and the corresponding VSE. All impulse responses were measured both with an omni-directional measurement microphone B&K 4192 and a B&K 4100 head and torso simulator (HATS) at the listening position. Impulse responses were measured for 32 positions with the same Dynaudio BM-6P loudspeaker that was used as the speech target source in the listening experiments. For the evaluation, the results were averaged over the 25 source positions for which the measurement distance was 2 m or larger.

2.2.2. Hearing aid directivity

Deviations of the auralized sound field from the original one were assumed to decrease the efficiency of the BF, which relies on the input from the two microphones, and, in turn, to decrease speech intelligibility. To assess the directional characteristics of the HAs, transfer functions were measured with the HA used in the test on the right ear of a B&K 4128 HATS. This was done for all incidence angles in steps of 10° at a distance of 2 m in an anechoic chamber, in the classroom, and in the VSE with each rendering method. All transfer functions were computed relative to the response of the HA in the omnidirectional program, measured on a B&K 4157 ear simulator with an outer-ear simulator DB 2012 for 0° incidence angle in an anechoic chamber. To reduce the strong magnitude fluctuations in the room transfer functions, their magnitude was smoothed with a 1/3-octave wide moving average filter.

2.3. Perceptual evaluation

2.3.1. Listeners

Eight normal-hearing native Danish speaking listeners (6 male, 2 female) with an average age of 27 years participated in the study and were paid an hourly wage. They were given written as well as oral information about the experiment and signed a consent form. The experiment was approved by the Danish Science-Ethics Committee (reference H-3-2013-004). The listeners were instructed in the use of the HAs as to changing the program and inserting or taking out the HAs after instruction. They were supplied with regular production receiver-in-the-ear Oticon Ino HAs providing a linear gain of 15 dB across the frequency range of the HA. In the HAs, an omnidirectional microphone and a static beamformer program could be selected. The HAs were coupled to the ears with mushroom-shaped silicone Oticon power domes,

such that no individual earmoulds were needed. All adaptive features of the HAs, like noise reduction and feedback cancellation, were turned off.

2.3.2. Stimuli

SRTs were measured using the Danish Dantale II speech-in-noise test [35], the Danish version of the Swedish Hagerman test [36]. This speech corpus is a matrix test spoken by a female talker that consists of 160 five-word sentences with an identical syntax of "name + verb + numeral + adjective + object". All sentences are permutations of the 50 words of a base list with 10 sentences, which makes the sentences hard to memorize and allows for reusing them within the same test session [37]. The masking noise was the corresponding Dantale II speech-shaped noise, produced from the test sentences that were superimposed with random pause durations for each sentence [35]. The target speech was embedded in clips of the noise file with a random start sample, such that the noise started 0.9 s before the sentence onset and ended 0.5 s after the end of the sentence. The on-and offset of the noise was windowed with 200 ms hanning ramps.

2.3.3. Experimental procedure

Before the actual measurements, the listeners were trained with 80 sentences, both with and without HAs and with both HA programs. The test conditions were counterbalanced across all listeners and the sentence lists were randomized with the constraint that no list could be re-used within seven runs. For each test condition, the SRT, representing the SNR at which 50% of the words were understood correctly, was determined in an adaptive procedure using two lists, i.e., 20 sentences. The level of the speech-shaped noise was kept constant at 70 dB SPL in all unaided conditions, and 62 dB SPL in all HA conditions, resulting in roughly equal loudness across the two conditions. The speech level was adjusted using an adaptive maximum-likelihood procedure [38]. The test was conducted in the patient-based, closed-set version [39], where the listener had to choose the correct words from all possible alternatives in a Matlab-GUI on an iPad. The target speech source was placed at 0° at distances of 2 m and 5 m, respectively, as shown in Figure 2. Three noise sources were placed at angles of ±112.5° and 180° at a fixed distance of 2 m. All loudspeakers were placed with their acoustic centre at ear level, i.e., about 120 cm above the ground.

An overview over the test conditions can be found in Table I. All listeners were tested in the classroom and in the VSE with both NLS and HOA rendering for the target distances of 2 m and 5 m. This was done without HAs as well as with the two HA programs. Half of the participants were first tested in the VSE, the other half of the participants was first tested in the classroom. During the SRT measurement, the listeners were asked to sketch the perceived position and extent of the sound sources in each experimental run on a response sheet with a schematic drawing of the listening test setup. The listeners were encouraged to describe any peculiarities they observed orally to the experimenter. Even though no formal evaluation was performed on these responses, the descriptions were expected to provide some hints regarding potential weaknesses of the auralization procedure or to allow for some exploration in the case of unexpected results. The experiments were divided into two sessions of about two hours.

Room	Distance	HA
R019	2 m	Unaided
VSE-NLS	5 m	Omni
VSE-HOA		BF

Table I: Overview over listening test conditions. All listeners performed the experiments in all combinations of the listed conditions.

3. Results

3.1. Physical evaluation

3.1.1. Room acoustic parameters

Figure 3 shows T_{30} (left panel) and C_{50} (right panel) measured in the classroom (square symbols) and in the VSE using NLS (crosses) and HOA rendering (circles). The symbols indicate the average values measured at the listening position shown in Figure 2 for the 25 source positions with a minimum distance of 2 m. The average value of T_{30}, determined as the average of the values for the 500 Hz and 1 kHz octave bands according to [11], was 0.49 s in the classroom and 0.53 s in the VSE with both rendering methods. The values in the classroom varied between 0.48 s at 1 kHz and 0.6 s at 2 kHz and dropped to 0.44 s at 8 kHz. In the lowest two frequency bands, no meaningful values could be determined in the classroom due to distinct room modes. Considering the limited frequency range of hearing aids, these frequency bands were not considered crucial and the values were omitted in the figure. The ODEON simulation results for T_{30} were essentially identical with the ones measured in the VSE, and thus omitted in the figure for clarity. This indicates that the reverberation time is well-preserved by the LoRA processing and that the playback room does not provide additional reverberation, which is in good agreement with [10], where similar measures were computed from the multichannel RIR. The values measured in the VSE differ from the ones in the classroom by

less than 0.1 s. This deviation corresponds to the calibration error of the ODEON model. An even closer match between room model and reality would have required the use of materials that are highly absorbent in very narrow frequency bands, which would have compromised the plausibility of the room model.

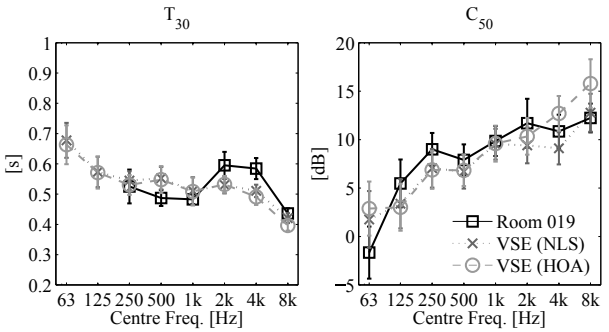


Figure 3: Average reverberation time T_{30} and clarity for speech C_{50} at the listening position for 25 source positions. The values were measured in the real classroom (square symbols) and in the VSE (crosses and circles).

Since the clarity for speech C_{50} represents the ratio of acoustic energy between the first 50 ms and the remaining part of the impulse response, it shows the opposite trend compared to the reverberation time. Apart from the two lowest frequency bands, the values ranged from 8 dB to 12.2 dB in the classroom. The values in the VSE tended to be slightly lower with a maximum deviation of 2.3 dB at 2 kHz. Bradley and colleagues [13] argued that a just noticeable difference of 3 dB for clarity represents a realistic value in real listening situations. Thus, the match between the room acoustic simulation and the real room may be sufficient for a convincing auralization. However, in the 125 Hz frequency band, the values measured in the VSE are about 5 dB lower than the simulated values obtained with ODEON. This difference is most likely caused by the playback room, which is not fully anechoic and produces some reflections in this frequency band. At the highest two frequencies, the clarity values for the HOA rendering method are markedly higher than the ones for NLS. Favrot and Buchholz [10] found a similar trend for the microphone position in the centre of the loudspeaker array. They explained this deviation by the energy regularization decoding method that is used in the frequency bands above the upper frequency limit imposed by the limited number of loudspeakers with HOA to preserve the total energy in the sweet spot.

Figure 4 shows the $IACC$ measured at the listening position in the classroom (square symbols), the VSE using NLS (crosses) and HOA coding (circles), for the two target source positions at 2 m (left panel) and 5 m (right panel) as a function of frequency. Two main trends can be observed: First, the $IACC$ for the 5-m target distance is lower than the corresponding value for the 2-m distance in nearly all room conditions. Second, in most cases, the IACC measured in the classroom is higher than in the VSE. Lower coherence values for larger distances were expected, because the sound field in a room becomes increasingly dominated by the reverberant sound with increasing distance. The lower values found in the VSE compared to the classroom may reflect the spatial 'jitter' introduced by the NLS technique and the imperfect reproduction of the sound field at the two ears with HOA coding. The pronounced dip in the curves at 500 Hz coincides with the decoupling frequency described by Lindevald and Benade [40]. They stated that the spatial average of the correlation function between the two ear signals in a room is well described by a modified sinc function with the first zero at about 500 Hz, representing the decoupling frequency. Below this frequency, the signals at the two ears are highly correlated, whereas above it, the signals are essentially two independent samples of the sound field. Lower $IACC$ values in the VSE might indicate a more diffuse sound field than in the real room, which would make a BF algorithm less effective.

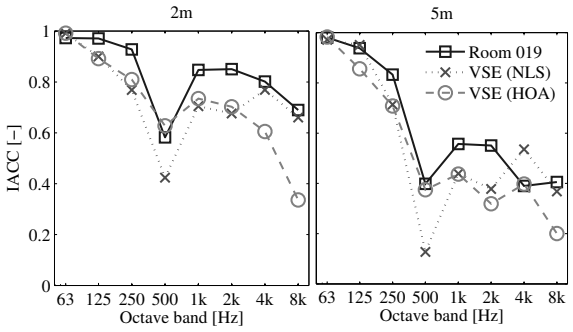


Figure 4: Interaural cross-correlation coefficient (IACC) measured in the real room (squares) and the VSE with NLS (crosses) and HOA rendering (circles) at a target source distance of 2 m (left panel) and 5 m (right panel).

3.1.2. Hearing aid directivity

Figure 5 shows the directivity patterns measured for the HA in the anechoic chamber (upper panels), Room 019 (middle panels), and the VSE with HOA rendering (bottom panels). The left column shows the directivity pattern for the omnidirectional program, the right column shows the pattern for the BF program. In the anechoic chamber (top row), the head

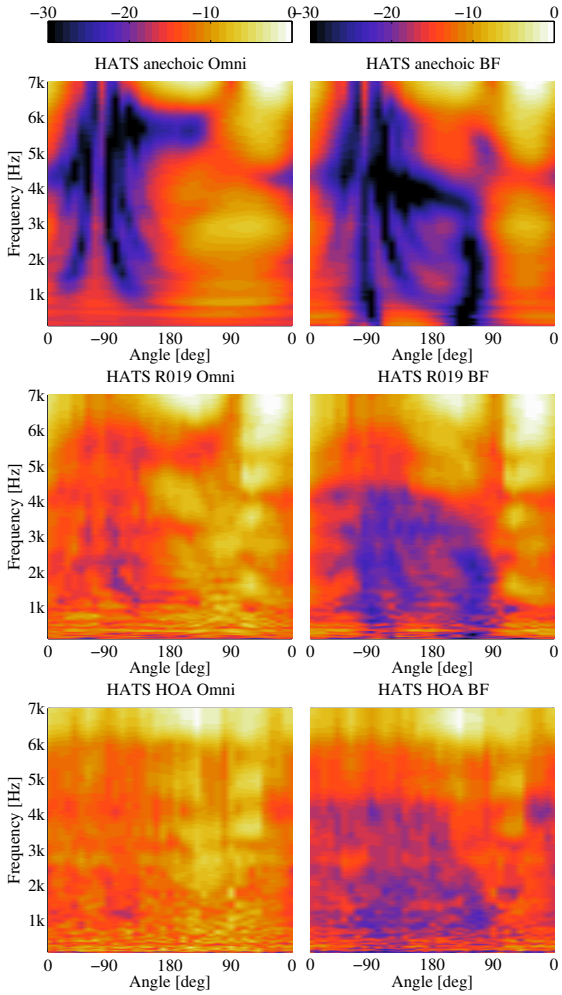


Figure 5: Directivity patterns of the HA measured on the right ear of a B&K HATS 4128 in an anechoic chamber (top row), the classroom (middle) and the VSE (bottom row). The left column shows the results for the omnidirectional program, the right column shows the results for the beamformer. All transfer functions are computed relative to the Omni-program for frontal (0°) incidence measured on an ear simulator B&K 4157 under anechoic conditions.

shadow and the interference patterns on the contralateral side of the head are clearly visible as dark areas. In addition, the BF results clearly show the zeros of the BF at about −100° and +120°, especially at the lower frequencies up to about 2 kHz. In Room 019 (right middle panel), remainders of the pattern can still be found, but the dynamic range between the highest and the lowest sensitivity is strongly reduced. This was expected since, unlike in an anechoic chamber where all the sound energy arrives from the direction of the source, the sound that arrives at the HA in a room also contains reflected energy from the different surfaces, which makes the sound field more diffuse. Even if a zero in the BF sensitivity pattern would perfectly eliminate the direct sound, e.g., generated from a noise source in the room, the microphone would still pick up most of the reflected sound. Using HOA rendering of the VSE, the dynamic range is further reduced, especially when comparing the values for a given frequency across the different incidence angles, i.e., values lying on a horizontal line in the plots. The zeros at the low frequencies can hardly be observed anymore. This indicates that the sound field inside the VSE might be even more diffuse than the one in Room 019. The results for NLS coding are not shown here because they are very similar to the results obtained for HOA.

3.2. Speech intelligibility

Figure 6 shows the mean value and standard deviation of the measured SRTs for the conditions listed in Table I, i.e., the three HA conditions 'unaided' (UA), 'Omni', and 'BF' measured in the three room conditions 'R019', 'VSE-NLS' and 'VSE-HOA' for target source distances of 2 m and 5 m. For the target source distance of 2 m (black symbols), the SRTs for the unaided conditions were found at -13.8 dB in the real room (R019, left panel), -11.8 dB in the VSE with NLS coding (middle panel), and -9.4 dB with HOA coding (right panel). The higher SRTs obtained with HOA compared to NLS coding are consistent with findings in an earlier study [14]. Using HAs in the omnidirectional microphone setting generally increased the average SRT compared to the unaided condition by up to 4 dB in the real room, whereas using HAs in the BF program lowered it by up to 2.7 dB with HOA coding. For the target source distance of 5 m (grey symbols) in Room 019 (left panel), the listeners showed an increase in SRT of about 3 dB in all HA conditions compared to the results obtained at 2 m. This was expected since the direct-to-reverberant sound ratio in a room usually decreases with increasing distance, which is generally assumed to have an adverse effect on speech intelligibility [41]. Compared to the results for the 2-m distance, the SRTs measured for the 5-m distance showed a considerably larger spread in the real room. At this distance, small head movements subjectively had a larger effect on the SRT than at 2 m and some listeners might have utilized them more successfully than others. This might be due to wave phenomena like standing waves and local interference patterns. This would also explain, why this effect is not seen in the VSE, because the ODEON model is based on geometrical acoustics and hence cannot capture wave phenomena.

For statistical analysis, a linear mixed model was fitted to the data with 'Room', 'Distance', and 'HA condition' as fixed factors and 'Listener' as random factor. In an Analysis of Variance (ANOVA), all factors and all two-factor interactions showed significant effects, indicating that there are differences

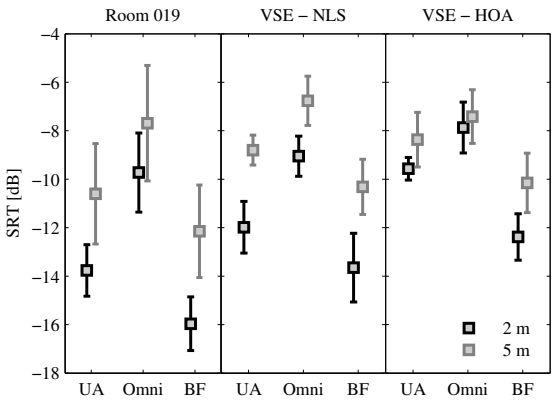


Figure 6: Average SRTs measured in Room 019, the VSE with NLS rendering and the VSE with HOA rendering for each of the HA conditions Unaided (UA), Omni, and Beamformer (BF), and for a distance of 2 m (black symbols) and 5 m (grey symbols). The error bars indicate ± one standard deviation.

between the results measured in the classroom and in the VSE. When only the data from Room 019 were considered, only the two main effects 'Distance' and 'HA condition' were significant, whereas their interaction was not. To address which VSE rendering method yields results that are more comparable to the real room, two ANOVAs were performed to compare the results of each rendering method to the ones measured in Room 019. In both cases, all main effects were highly significant, including the factor 'Room', which indicates that the measured SRTs measured in the room are different from the ones in the classroom. However, all two-factor interactions showed significant effects in the case of HOA rendering, but not in the case of NLS rendering ($\alpha = 0.05$). Especially the difference in SRT between the two distances with NLS (Figure 6, middle panel) was found to be similar as in Room 019 (left panel), whereas the pattern looks clearly different for HOA (right panel). This is reflected in a non-significant interaction between 'Room' and 'Distance' [$F(1,79) = 0.1441$, $p = 0.7053$] with NLS, whereas the same interaction was significant with HOA [$F(1,79) = 9.9380$, $p = 0.0023$]. This suggests that NLS coding preserves more of the cues that contribute to speech intelligibility, despite the simple algorithm, especially with respect to distance.

Since a VSE system will probably mostly be used to compare perceptual outcome measures in different conditions, the benefit in SRT from the BF over the omnidirectional program was computed as $SRT_{\mathrm{Omni}} - SRT_{\mathrm{BF}}$ (cf., Figure 7). In Room 019, this benefit was, on average, 6.2 dB for a target distance of 2 m, while it dropped to about 4.5 dB for the 5-m distance. The values measured in the VSE were found to be slightly lower in all cases. With NLS, the values dropped to 4.6 dB at 2 m distance, and to 3.5 dB for the 5-m distance. With HOA, the average benefit was 4.3 dB for the 2-m distance and 2.9 dB for the 5-m distance. An ANOVA on these benefits again showed significant main effects of the factors 'Room' and 'Distance', indicating that the BF benefit is not equal, but smaller in the VSE than in the real room, and decreases with increased distance. However, a set of one-sample t-tests showed that the mean value underlying the measured benefits was larger than zero in all conditions, indicating that the BF yielded a clear advantage in speech intelligibility relative to the omnidirectional processing in all tested conditions.

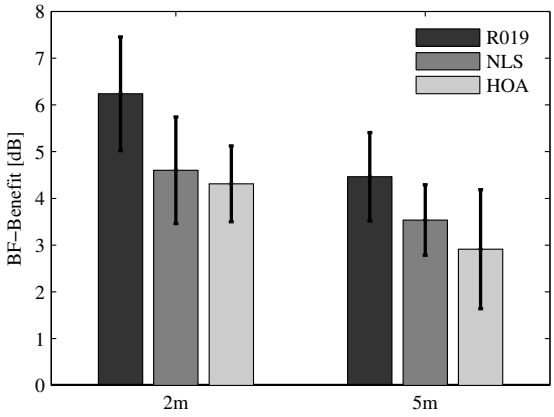


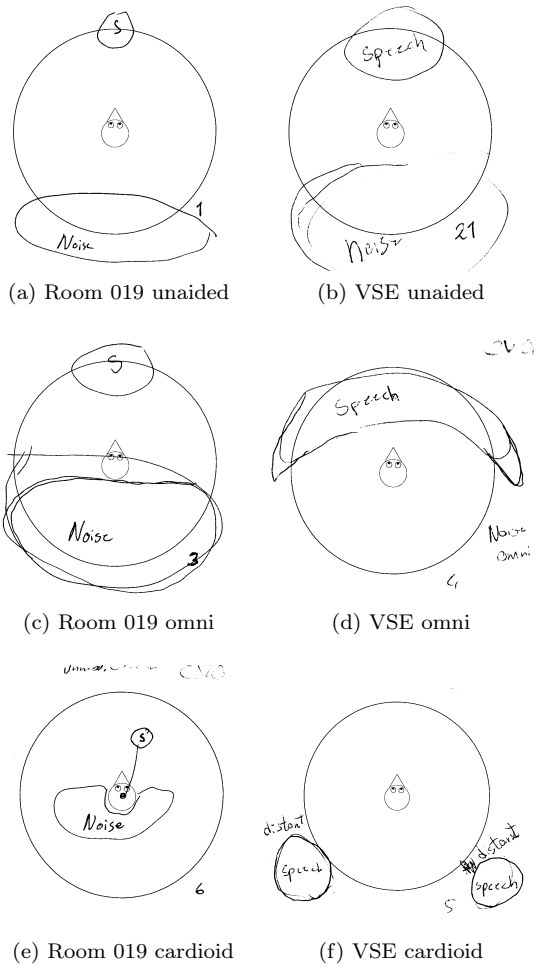
Figure 7: Benefit from the BF algorithm over the omnidirectional microphone pattern for all room conditions and the two target source distances. Higher values indicate better performance, the error bars indicate ± one standard deviation.

3.3. Subjective impression

In each run, the listeners were also asked to sketch their subjective impression of localization and extent of the sound sources in a schematic drawing of the listening situation with a listener and a circle indicating the radius of the loudspeakers. In the real room, the result tended to change from a very clear and focused image in the unaided case (see Figure 8a for an example) to a spatially much less defined image with HAs in the omnidirectional setting (Figure 8c). This impression may have resulted from the loss of the directional-dependent pinna cues due to the microphone position above the ear. Switching to the BF program, many of the listeners again reported a change in the spatial impression. Often, the sound sources were described as being closer around the head and sometimes the target speech was perceived inside the head, i.e., internalized (Figure 8e). Some listeners also reported hearing the noise source inside the head, while the speech was located outside. In the VSE, the virtual

sound sources were often perceived as being wider and less well-defined than in the classroom (Figure 8b). Especially the three noise sources were often fused into a single percept or the listeners reported that the noise was 'somewhere behind' them; some listeners described the speech as sounding more reverberant. The noise sources were perceived even wider when the HAs were used with the omnidirectional program. In this setting, many listeners perceived the noise as coming from all around the room. The speech source was often described as being much broader than in the classroom (Figure 8d). With the BF program, the descriptions became more diverse. Some listeners again reported the target speech to be closer to them or even inside their head, in some cases the sound image split and was indicated at different places (Figure 8f). The noise sources were often perceived at two separate locations, either close to the ears or at loudspeaker distance at the sides of the array. Even though there was a lot of variability in the subjective impression, it was clear that all conditions with hearing aids tended to distort the spatial perception of direction, source width, and distance. Interestingly, some listeners had the impression that they performed much worse in the BF than in the Omni conditions, even though their SRTs were actually consistently better.

Finally, some listeners reported that the transition from understanding the whole sentence to not understanding anything seemed less gradual in the VSE than in the classroom, which is reflected in the generally smaller variability in the data obtained in the VSE compared to the real room. This might indicate that the underlying psychometric function is actually steeper in the VSE than in the real room, which would imply that the sensitivity of the speech test is actually higher inside the VSE.



(a) Room 019 unaided (b) VSE unaided

(c) Room 019 omni (d) VSE omni

(e) Room 019 cardioid (f) VSE cardioid

Figure 8: Subjective evaluation of listening test conditions. The scans show the descriptions of one listener in Room 019 (left) and the VSE (right) for the Unaided condition (upper), the Omni program (middle) and the BF (bottom), respectively. In conditions (d) and (f) the listener indicated that the noise was perceived as coming from all directions.

4. Discussion

4.1. Physical evaluation

The results from the physical measurements should provide some insights regarding the different limiting factors in the auralization chain: the ODEON simulation, the auralization system with the LoRA toolbox and the loudspeaker array, and the playback room. A room acoustic computer model can only provide a rough approximation of the actual sound field in a room. Inside such a model, the room geometry needs to be simplified and usually assumptions need to be made regarding the materials in the room and their acoustical properties. Typically, room acoustic simulation programs are evaluated in terms of their prediction of room acoustic parameters, e.g., [42]. Here, the measured room acoustic parameters agreed well between the ODEON simulation and the real room. The values for T_{30} and C_{50} measured in the VSE agreed very well with the ODEON results, indicating that the temporal energy decay in the playback room closely follows the model and that the playback room is sufficiently dampened. Lower values for the $IACC$, however, indicated that there are differences in the spatial characteristics of the sound field between the real room and the VSE and that the sound field reproduced inside the loudspeaker array is more diffuse than the one in the classroom. This might, at least partly, account for the larger perceived spaciousness and reverberance. Another indication of a more diffuse sound field in the VSE is the reduced directivity obtained with the BF algorithm in the HAs. The main source of the increased diffuseness is probably the finite number of loudspeakers, which imposes the limitation of a spatial quantization with the NLS method and the requirement to truncate the HOA

series after the 4th order which, in turn, limits the spatial resolution of the system. However, the usual room acoustic parameters might not be sufficient to describe the performance of the room acoustic models and the input data for the auralization system might also be a limiting factor for the authenticity of the VSE.

4.2. Listening experiments

In general, the VSEs could reproduce the trends in the SRT variations found in the real room very well, even though the SRTs were generally shifted towards slightly higher levels, indicating poorer speech intelligibility in the VSE. This finding is not surprising, because each step in the generation of the VSE, i.e., the ODEON simulation, the LoRA toolbox, and the loudspeaker array and playback room, imposes some limitations on the overall result. Most geometrical room acoustic simulation methods are only appropriate when the dimensions of the room are long compared to the wavelength [6] and therefore not very reliable at frequencies below the Schroeder frequency [43]. Another aspect that potentially limits the performance of the auralization system is the rendering method. If HOA is used, the number of the loudspeakers limits the Ambisonics order which, in turn, limits the localization accuracy. It also implies an upper frequency limit for correct sound field reproduction. In the system under test, this frequency limit is at about 2.2 kHz if a sweet spot of 20 cm diameter is considered [10]. Above this frequency, the magnitude of the sound is still correct, but the phase relations might be incorrect. If the NLS technique is used instead, these limitations do not apply. However, in this case, the sound source positions are limited to the angles at which loudspeakers are available and the reflections are subject to spatial discretization, which might also blur the perceived localization of the sound source. If the localization accuracy is reduced compared to the real room, it might become more difficult to segregate the target speech from the noise leading to a higher SRT. If the playback room is not sufficiently close to anechoic, the natural reverberation will increase the reverberation in the VSE and will add a sense of increased spaciousness. In the system under test, however, this was not considered an issue due to the very short reverberation time.

Another result from this study was that the SRTs measured with HOA tended to be higher than the ones obtained with NLS. This finding is consistent with the results of an earlier study [14] that found higher intelligibility scores with NLS than with 4th order HOA which, in turn, were higher than the ones measured with 1st order Ambisonics. Differences between the SRTs measured with the two tested HA programs, however, could clearly be observed in all VSE conditions and they were similar to the ones measured in the classroom. This is an important finding since it demonstrates that the results measured in the realistic VSE seem to be a good indicator of real-world performance. Also for other differential measures, e.g., the comparison of the listening performance in several simulated rooms with different acoustical properties [16], the VSE seems to be well-suited.

Regarding the reports of the subjective impression of the perceived position and the extent of the sound sources, visual cues might have contributed to the result that the sound sources were usually perceived as wider in the VSE than in the classroom. The listeners were surrounded by 29 loudspeakers in the VSE, whereas there were only four single loudspeakers in the classroom. The role of potential visual cues in the evaluation cannot be clarified in the present study. However, in all experimental conditions, the sources in the VSE were simulated at angles at which there were loudspeakers in the array, which might have helped to consolidate the auditory image.

4.3. Perspectives

The auralizations in this study were based on room simulations. This approach has the major advantage that it makes the auralization method very flexible. Existing models can easily be adapted to new listening situations with, e.g., additional sound sources. Furthermore, it is possible to auralize rooms that do not physically exist (yet) or acoustic situations that do not occur in real rooms, but allow for the study of basic aspects of spatial hearing, e.g., the influence of single reflections on speech intelligibility [44]. One limitation, however, is that while the method works well for static scenes, it is quite cumbersome to implement moving sound sources. Furthermore, the inherent limitations of ray-tracing based room acoustic models do not allow accurate reproduction of low-frequency effects, like room modes, and only roughly represent the acoustic properties of a room. Also fast fluctuations in the reverberant tail of the room impulse response are difficult to capture with the present system.

Some limitations can be overcome when the auralization is based on array microphone recordings instead of room simulations. A recent study [45] used multiple VSEs in a loudspeaker array similar to the one used in the present study that were recorded with a spherical 32-microphone array and rendered using a direct inversion method. This method was shown to lead to a very convincing auralization of complex scenes, even with moving sources. However, this happens at the cost of reduced flexibility because the scene cannot be changed once recorded. A spherical HOA microphone array with 52 1/4-inch microphones in a rigid sphere with a diameter of 10 cm has been

developed and is currently being tested [46]. With this technique, array recordings of real acoustic scenes can be combined with simulation techniques to place target or interfering sources in a virtual scene. This could be done either by recording the background scene directly and by measuring impulse responses at the same position without background noise (which might not always be possible), or by combining the background recordings with target sources based on a room simulation.

5. Summary and conclusion

In this study, speech intelligibility in noise was used as a measure to assess the authenticity of a VSE based on a carefully calibrated room acoustic model of an existing classroom. The VSE was compared to the real room by means of T_{30}, C_{50}, and $IACC$. It was found that the average values for T_{30} and C_{50} measured in the VSE were very close to the values simulated in ODEON. The slight differences between the parameters measured in the classroom and the VSE were most likely caused by the setup of the room model in ODEON rather than by the LoRA processing or the reproduction room. However, the $IACC$ was found to be lower in the VSE than in the real room. The HA directivity patterns showed a reduced level of detail in the classroom compared to the anechoic chamber and a further reduction in detail in the VSE as a consequence of the slightly more diffuse sound field in the VSE compared to the real room.

In the listening experiments, the SRTs were generally found to be slightly higher in the VSE than in the classroom. It was shown that the SRTs in the VSE in the conditions with HAs improved when a static BF was used instead of an omnidirectional microphone, even though the improvement was slightly smaller than in the real room. Furthermore, the dependence of the SRT on the target source distance was found to be very similar in the VSE and in the classroom, when the NLS rendering method was used. The NLS method thus seems to preserve more of the crucial acoustical features of a real room than HOA.

Even though the SRTs differed between real room and simulation, all differential results translated well to the real world. Since the evaluation of new HA signal processing features typically considers such differential measures, the VSE system may represent a valuable tool for such testing where end users can be involved early in the HA development process. For the time being, NLS should be preferred over HOA for experiments in which the reduced spatial resolution of NLS compared to HOA is not too critical, like speech intelligibility experiments, because it seems to preserve more of the underlying cues.

Acknowledgements

The authors wish to thank the editor and the two anonymous reviewers for their constructive feedback and Sylvain Favrot and Pauli Minnaar for their valuable contribution to this study. This work was supported by a research consortium with Oticon, Widex and GN ReSound. Parts of this work were presented at the AIA-DAGA Conference on Acoustics in Merano, Italy, 18-21 March 2013.

References

[1] A. W. Bronkhorst: The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acta Acustica united with Acustica **86** (2000) 117–128.

[2] H. Møller: Fundamentals of binaural technology. Applied acoustics **36** (1992) 171–218.

[3] A. J. Berkhout, D. de Vries, P. Vogel: Acoustic control by wave field synthesis. The Journal of the Acoustical Society of America **93** (1993) 2764–2778.

[4] M. A. Gerzon: Periphony: With-height sound reproduction. Journal of the Audio Engineering Society **21** (1973) 2–10.

[5] J. Daniel, R. Nicol, S. Moreau: Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging. Preprints-Audio Engineering Society (2003).

[6] M. Vorländer, J. E. Summers: Auralization: Fundamentals of acoustics, modelling, simulation, algorithms, and acoustic virtual reality. 1. ed. Springer-Verlag, Berlin, 2008.

[7] V. Pulkki: Spatial sound reproduction with directional audio coding. Journal of the Audio Engineering Society **55** (2007) 503–516.

[8] B. Seeber, S. Kerber, E. Hafter: A system to simulate and reproduce audio-visual environments for spatial hearing research. Hearing research **260** (2010) 1–10.

[9] C. L. Christensen: ODEON Room Acoustics Software, Version 12, User Manual, ODEON A/S, Kgs. Lyngby, Denmark. 2013.

[10] S. Favrot, J. Buchholz: LoRA: A loudspeaker-based room auralization system. Acta Acustica united with Acustica **96** (2010) 364–375.

[11] EN ISO 3382-1: Acoustics – measurement of room acoustic parameters – part 1: Performance spaces. 2009.

[12] T. J. Cox, W. Davies, Y. W. Lam: The sensitivity of listeners to early sound field changes in auditoria. Acta Acustica united with Acustica **79** (1993) 27–41.

[13] J. S. Bradley, R. Reich, S. Norcross: A just noticeable difference in C_{50} for speech. Applied Acoustics **58** (1999) 99–108.

[14] S. Favrot, J. M. Buchholz: Validation of a loudspeaker-based room auralization system using speech intelligibility measures. Audio Engineering Society Convention 126, 2009, Audio Engineering Society.

[15] S. Favrot, J. Buchholz: Distance perception in loudspeaker-based room auralization. Proc. 127th AES Convention, 2009.

[16] P. Minnaar, C. Breitsprecher, M. Holmberg: Simulating complex listening environments in the laboratory for testing hearing aids. Proc. Forum Acusticum, 2011.

[17] V. Best, G. Keidser, J. M. Buchholz, K. Freeston: An examination of speech reception thresholds measured in a simulated reverberant cafeteria environment. International journal of audiology (2015) 1–9.

[18] G. Grimm, S. Ewert, V. Hohmann: Evaluation of spatial audio reproduction schemes for application in hearing aid research. Acta Acustica united with Acustica **101** (2015).

[19] C. Oreinos, J. M. Buchholz: Objective analysis of ambisonics for hearing aid applications: Effect of listener's head, room reverberation, and directional microphones. The Journal of the Acoustical Society of America **137** (2015) 3447–3465.

[20] W. Yang, M. Hodgson: Validation of the auralization technique: Comparative speech-intelligibility tests in real and virtual classrooms. Acta Acustica united with Acustica **93** (2007) 991–999.

[21] M. Hodgson, N. York, W. Yang, M. Bliss: Comparison of predicted, measured and auralized sound fields with respect to speech intelligibility in classrooms using catt-acoustic and odeon. Acta Acustica united with Acustica **94** (2008) 883–890.

[22] M. Rychtáriková, T. Bogaert, G. Vermeir, J. Wouters: Perceptual validation of virtual room acoustics: Sound localisation and speech understanding. Applied Acoustics **72** (2011) 196–204.

[23] M. Schoeffler, J. Gernert, M. Neumayer, S. Westphal, J. Herre: On the validity of virtual reality-based auditory experiments: a case study about ratings of the overall listening experience. Virtual Reality (2015) 1–20.

[24] M. Kleiner: Speech intelligibility in real and simulated sound fields. Acta Acustica united with Acustica **47** (1981) 55–71.

[25] T. Koski, V. Sivonen, V. Pulkki: Measuring speech intelligibility in noisy environments reproduced with parametric spatial audio. Audio Engineering Society Convention 135, 2013, Audio Engineering Society.

[26] H. Dillon: Hearing aids. Thieme Medical Pub, 2001.

[27] M. Valente, D. Fabry, L. G. Potts: Recognition of speech in noise with hearing aids using dual microphones. Journal of the American Academy of Audiology **6** (1995).

[28] J. Wouters, L. Litière, A. Van Wieringen: Speech intelligibility in noisy environments with one-and two-microphone hearing aids. International Journal of Audiology **38** (1999) 91–98.

[29] J. Rindel: The use of computer modeling in room acoustics. Journal of Vibroengineering **3** (2000) 41–72.

[30] J. Rindel, C. Christensen: Room acoustic simulation and auralization–how close can we get to the real room? WESPAC 8, The Eighth Western Pacific Acoustics Conference, Melbourne, April 2003, Citeseer.

[31] J. Blauert: Spatial hearing: the psychophysics of human sound localization. The MIT Press, 1997.

[32] R. Litovsky, H. Colburn, W. Yost, S. Guzman: The precedence effect. The Journal of the Acoustical Society of America **106** (1999) 1633–1654.

[33] J. Buchholz, J. Blauert, J. Mourjopoulos: Room masking: Understanding and modelling the masking of reflections in rooms. Audio Engineering Society Convention 110, 5 2001.

[34] S. Müller, P. Massarani: Transfer-function measurement with sweeps. Journal of the Audio Engineering Society **49** (2001) 443–471.

[35] K. Wagener, J. Josvassen, R. Ardenkjær: Design, optimization and evaluation of a Danish sentence test in noise. International Journal of Audiology **42** (2003) 10–17.

[36] B. Hagerman: Sentences for testing speech intelligibility in noise. Scandinavian Audiology **11** (1982) 79–87.

[37] K. C. Wagener, T. Brand: Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters. International Journal of Audiology **44** (2005) 144–156.

[38] T. Brand, B. Kollmeier: Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. The Journal of the Acoustical Society of America **111** (2002) 2801.

[39] E. Pedersen: Bestemmelse af taleforståelighed i støj [Danish]. Diploma Thesis. Syddansk Universitet, Odense, August 2007.

[40] I. Lindevald, A. Benade: Two-ear correlation in the statistical sound fields of rooms. The Journal of the Acoustical Society of America **80** (1986) 661–664.

[41] J. Bradley, H. Sato, M. Picard: On the importance of early reflections for speech in rooms. The Journal of the Acoustical Society of America **113** (2003) 3233.

[42] I. Bork: Report on the 3rd Round Robin on Room Acoustical Computer Simulation Part II: Calculations. Acta Acustica united with Acustica **91** (2005) 753–763.

[43] M. R. Schroeder, K. H. Kuttruff: On frequency response curves in rooms. comparison of experimental, theoretical, and monte carlo results for the average frequency spacing between maxima. The Journal of the Acoustical Society of America **34** (1962) 76–80.

[44] I. Arweiler, J. Buchholz: The influence of spectral characteristics of early reflections on speech intelligibility. Journal of the Acoustical Society of America **130** (2011) 996.

[45] P. Minnaar, S. F. Albeck, C. S. Simonsen, B. Søndersted, S. A. D. Oakley, J. Bennedbæk: Reproducing real-life listening situations in the laboratory for testing hearing aids. Audio Engineering Society Convention 135, 2013, Audio Engineering Society.

[46] M. Marschall, S. Favrot, J. Buchholz: Robustness of a mixed-order ambisonics microphone array for sound field reproduction. Audio Engineering Society Convention 132, 2012, Audio Engineering Society.